

EML - the Environmental Markup Language

Hans Knud Arndt, Thomas Bandholtz, Oliver Günther, Maria Rüter, Thomas Schütz

maria.ruether@uba.de; thomas.schuetz@uba.de (Federal Environmental Agency, Berlin);
arndt@wiwi.hu-berlin.de; guenther@wiwi.hu-berlin.de (Humboldt University, Berlin);
thomas.bandholtz@sema.de (Sema Group, Cologne).

Abstract.

In this paper, we propose a first design for an XML-based Environmental Markup Language. The proposal builds heavily on existing de facto standards for environmental metadata, such as CDS, UDK, and GEIN. EML serves to wrap and, wherever possible, harmonize these standards. EML consists of two parts: EML MetaData and EML Data eXchange. EML MetaData is a special namespace and vocabulary used by information brokers that are part of the *Semantic Web* and have a special focus on environmental concerns. EML Data eXchange consists of a core set of attribute and document type definitions that shall be used when exchanging any kind of environmental data content.

1 Who needs EML?

XML, the extensible Markup Language, is rapidly developing to become the *lingua franca* of the internet. In this paper, we discuss the significance of these developments for environmental data management and propose a first design for an XML-based Environmental Markup Language (EML).

1.1 Archives of observation data

Over the past 100 years, the observation and measurement of environmental phenomena in the air, the water, the soil, the food, and the atmosphere has yielded enormous amounts of raw data and derived summary data.

Only a small percentage of these terabytes of data are stored under productive conditions, i.e., in direct access of a working information system. The large majority of these data sets are no longer usable or, at the very least, no longer used. They are stored on outdated storage media or in outdated data formats; retrieving the information would require a considerable amount of manual labor and thus be very costly. Moreover, many of the IT pioneers who started these developments are now close or past requirement age, taking with them most of the knowledge about the structure and quality of those data sets. Tape readers and software that would be necessary to access this knowledge base no longer work or cannot be re-installed as they require antique hardware out of use. Thus the mass of environmental data cannot be accessed, and even if it could be, nobody would be able to understand their semantics.

What if all that data would exist in a common, well structured format, accessible by almost any kind of computer system, and understandable because of its self-contained explicit data-structure and semantic assertions?

1.2 Operative integration of environmental data

Environmental Management Information Systems (EMIS) support the management of environmental information, in particular data about material and energy flows within a company. Crucial data about material and energy flows and their impact on the environment can not only be found within the borders of the company, but often also reside with suppliers and companies that process outputs further or handle the disposal of waste. The integration of data is

necessary for a meaningful analysis of environmental effects of production. The heterogeneity of data, however, may impede this integration considerably. Different hardware and software platforms, incompatible data formats and standards, as well as insufficient documentation with meta-information are the reasons.

Metadata could pragmatically be described as "data about datasets", including data about the representation, the management, the quality, and the content reference of environmental objects.

Just imagine there were a standard about which meta-information had to be collected and maintained along with data about material and energy flows. What if there was a common approach to data integration focusing on the transmission of information from various source systems? XML provides a solution to these problems.

1.3 Information brokerage

Early information brokers, like GELOS, GILS, CDS, JRC, EIONET should be well known to the ISESS community. All of them are based on pre-internet standards like Z39.50 or SGML. More recently, the German Federal Environmental Agency has presented the *German Environmental Information Network* (GEIN)¹, an environmental information broker based on new-generation Internet XML technology, implemented by Sema Group. The data is stored in Software AG's Tamino XML server. The agency has established this countrywide information network as a resource for sharing its experience and knowledge on a national and international level, and supporting its active involvement in the promotion of environmental protection.

The dynamic development approach means that searches are no longer limited to simple static Web pages. In fact, it is possible to integrate information derived dynamically from databases via simple HTTP-Services. The overall protocol in GEIN is XML.

2 How EML uses XML

Everybody knows various kinds of textual markup. From this point of view, the only benefit of XML seems to be readability. This chapter wants to show some of the more sophisticated aspects of XML.

In an XML format, the beginning of this chapter might look like:

```
<title> How EML uses XML </title>
```

```
<abstract> Everybody knows various kinds of textual markup. From this point of view, the
only benefit of XML seems to be readability. This chapter wants to show some of the
more sophisticated aspects of XML. </abstract>
```

Nobody has any difficulty to read and understand this little example of an XML document. XML could also have expressed:

```
<title xml:lang="en"> How EML uses XML </title>, or
<title xml:lang="de"> Wie EML XML verwendet </title>
```

So far, XML is: Entities (<title>) + Attributes (xml:lang) that describe content (the *text* between start-tag and end-tag).

In a next step, you can use XML to serialize datasets that are highly structured and have nothing in common with text.

```
<air_immission cas_nr="4711" location="Berlin" date="2000-05-30T12:00" val="0.07" />
```

might be a (simplified) single dataset out of a measuring network sequence.

The names and the meaning of the entities and attributes are not predefined by XML, but agreed to by a specific community concerned with a particular application. That's why XML is called "extensible". In fact, almost every usage is an extension.

XML offers *Document Type* and *Schema Definitions* that allow parsers to validate documents by simply parsing the definition first. While DTD are a heritage from SGML, *schemas* refer to a more refined method of semantic modeling that introduces an object-oriented approach to the markup world.

XML *namespaces* are the solution to a very simple but ugly problem: if one community defines its own entities and attributes, there will often be another community using the same names with a different meaning.

The "GEIN 2000"-namespace (*g2k*, see <http://www.gein.de/2000/profile-11.htm>) is a working example. *g2k* defines tags like `<g2k:topic>`, which would be equivocal in the XML-world without the prefix *g2k*. Figure 1 shows an example of *g2k* metadata describing a document which is identified by an URI (Unified Resource Identifier) – in this case a web page.

```
<?xml version='1.0' encoding="iso-8859-1"?>
<g2k:G2K xmlns:g2k="http://www.gein.de/2000/profile-11#" >
  <g2k:description uri="http://www.kingfisher.de/doc.htm"
    xml:lang="en" date="2000-02-02">
    <g2k:portal>NL</g2k:portal>
    <g2k:class>NL10</g2k:class>
    <g2k:title>The King Fisher</g2k:title>
    <g2k:abstract> About his habits and reservates.
  </g2k:abstract>
    <g2k:topic uid="28753" term="kingfisher" rank="10"/>
    <g2k:topic uid="16963" term="bird species" rank="5"/>
    <g2k:area uid="1100000000" term="Berlin" type="town" rank="8"/>
    <g2k:area uid="NR-12" term="Havelland" type="reservat" rank="7"/>
    <g2k:time event="cal/2000" from="1999-12-31" to="2000-01-01"/>
  </g2k:description>
</g2k:G2K>
```

Figure 1 Example of a resource description using the *g2k* namespace.

The same "profile" contains an example that makes use of two namespaces, here "*g2k*" and "*dc*" (Dublin Core, Figure 2).

In the same way, EML can integrate existing namespaces like those of GELOS, GILS, CDS, JRC, or EIONET, if each of those would define their de facto namespace by means of XML. This is not a real harmonization, but it is already progress if things have been *integrated* if they cannot be *harmonized*. In any case: do they really have to be harmonized? There are different domain-specific interests behind each of these namespaces, and too much of harmonization would also lead to a certain loss of focus of these individual efforts. As Figure 2 shows in a very simple case, XML can be used with integrated namespaces, because each of them remains explicit.

```
<?xml version="1.0" encoding="iso-8859-1"?>
<g2k:G2K xmlns:g2k="http://www.gein.de/2000/profile-11#"
  xmlns:dc="http://purl.org/dc/elements/1.1/" >
  <g2k:specialQuery scope="Literature"
    xml:lang="en" mid="4711" match="and">
    <dc:Creator>Dr. B. King-Fisher</dc:Creator>
    <dc:Subject>Bird species, Kingfisher</dc:Subject>
    <g2k:topic uid="7827" term="Kingfisher" />
  </g2k:specialQuery>
</g2k:G2K>
```

Figure 2 Mixing two namespaces (*g2k* and *dc*) in one description

In the first version of its profile, GEIN tried to be compliant with the very restricted rules of the Resource Description Framework (RDF), the W3C's recommended metadata layer based on and written in XML. RDF only allows a "triplet" form of semantic assignment, with the intention that any general RDF machine in the world can immediately handle any Resource Description conform to RDF. In the end, it was decided that GEIN should use unrestricted XML for the internal network semantics and communication, but it still supports an external RDF interface.

What seems to be more important in this context is the use of a well-defined *vocabulary*. Having defined the tags, GEIN proceeded to using thesauri and classifications in the "tagged" content. GEIN uses a complex thesaurus comprising more than 50,000 environmental terms and another 50,000 geographic names. There is an environmental calendar function that associates dates of important events with common "nicknames" (i.e. "since Chernobyl"). In Figure 1, the attributes "uid" (<g2k:topic> and <g2k:area>) and "event" (<g2k:time>) represent references to identifiers of these thesauri. This is the only way to handle homonyms, synonyms and dialects in metadata.

Besides taking advantage of these features of GEIN, EML should make similar use of other thesauri, such as the "General Multilingual Environmental Thesaurus" (GEMET) maintained by the European Environmental Agency (EEA) in Copenhagen. GEMET is based on an integration of different national thesauri in Europe. In the current version 2.0 it contains "EnVoc" of UNEP - Infoterra and will probably be integrated in USA by EPA. GEMET contains 5,300 descriptors, 1,260 synonyms, and a complete glossary in 12 languages (English and most Western European languages).

Thesauri such as GEIN or GEMET can be stored in XML themselves (as proven by GEIN). For a final, international data model, ISO/IEC 13250 ("Topic Maps") should be considered.

A "natural" extension of the EML approach concerns the integration and documentation of models and methods. Typical examples include environmental simulation models or statistical aggregation techniques. In order to make such methods available to all users, one needs to carefully document them. Potential users need to know what a method does exactly and how to execute it. This includes information about its physical location and its input and output requirements. In a related project called MMM (Middleware for Method Management)² some of the authors have developed an XML-based approach to perform this task. Figure 3 shows an example of an XML record describing a Matlab module for time series analysis.

```
<?xml version='1.0' encoding="iso-8859-1"?>
<OBJECT> ...
  <ACCESS> location="extern" decompression="none" tarfile="arfit.tar.gz">
    http://www.aos.princeton.edu/WWWPUBLIC/tapio/arfit/adjph.m
  </ACCESS>
  <SIGNATURE>
    <CALL>adjph(x)</CALL>
    <INPUT>
      <PARAMETER>
        <ID>x</ID>
        <DESCRIPTION>complex matrix</DESCRIPTION>
      </PARAMETER>
    </INPUT>
    <OUTPUT> ... </OUTPUT>
  </SIGNATURE>
  <ENVIRONMENT>
    <ENGINE><MATLAB/></ENGINE>
  </ENVIRONMENT>
</OBJECT>
```

Figure 3 Description of a method interface (detail)

Of course, all these general-named entities (<object>!) can only be valid in a specific domain, whose namespace is implicit in the sample given in Figure 3.

3 A Short Definition of the EML Approach

XML (eXtensible Markup Language) has been developed by the World Wide Web Consortium (W3C). XML version 1.0 was published in spring 1998. Because of its platform independency and its extensive standardization, XML is qualified for the exchange of information and meta-information between heterogeneous information systems of organizations of all kinds. Increasingly, there are branch and application specific implementations of XML (e.g., the Chemical Markup Language CML or the Mathematical Markup Language MathML). Some of them already play an important role in the day-to-day business practice. However, there has been no existing XML initiative for the environmental domain yet.

EML (Environmental Markup Language) is a set of recommendations for the national and international usage of XML (eXtensible Markup Language) in the communication of environmental information.

EML consists of two parts:

- EML MetaData
- EML Data eXchange

EML MetaData is a special namespace and vocabulary used by information brokers that are part of the *Semantic Web* and have a special focus on environmental concerns. Any kind of information in the WWW that considers itself as a contribution to the global knowledge about the environment, should index itself with EML MetaData.

EML Data eXchange consists of a core set of attribute and document type definitions that shall be used when exchanging any kind of environmental data other than metadata. Both attribute and document types may be extended or overwritten for the needs of special sub-domains.

In the future, we expect a set of tools to be recommended and maintained by an agency that takes responsibility for EML.

3.1 EML MetaData

Environmental metadata has been established systematically since the early 90s of the last century after a decade of gathering environmental observation and measuring data. In a global, and even in a national view, all these data collections had been hard to find by new applications, and even harder to access or exchange. This was the age of *data catalogues* like GILS, GELOS, CDS, or the German UDK.

3.1.1 Data Catalogues

The original concept of data catalogues is based on well structured descriptions, separated from the data itself. Each data description points to an address of a person or institution that may be contacted to obtain the data itself.

One of the first systematic environmental data catalogue was the German "Umweltdatenkatalog"³.

In 1995 the European Environmental Agency established the European Topic Centre on Catalogue of Data Sources (ETC/CDS)

".. to develop a European Catalogue of Data Sources and a European Multilingual Environmental Thesaurus to provide access to environmental information on a European scale."⁴

Both efforts recently have been adopted by the United Nations Environment Program globally⁵. These efforts can be considered as useful and successful, but they are continuously limited by the need to gather explicit input of metadata by the cooperating providers. In the meantime there exists a comfortable interactive tool, but still it is difficult to collect complete and qualified data, and to keep it up-to-date.

In the later 90s, these catalogues integrated SGML interfaces, which now migrate to XML⁶. While this is a first step to conform with the EML proposal, there still remain different attribute sets implicitly. As ETC/CDS states⁷, the CDS dataset contains the GELOS dataset as well as parts of GILS and parts of Dublin Core, but this is not made explicit by using namespaces. The CDS DTD does not state, which of its elements are GELOS, GILS, or Dublin Core. As these have their own DTDs, we have redundant element definitions in these overlapping metadata sets. This is likely to cause misunderstandings and ambiguities.

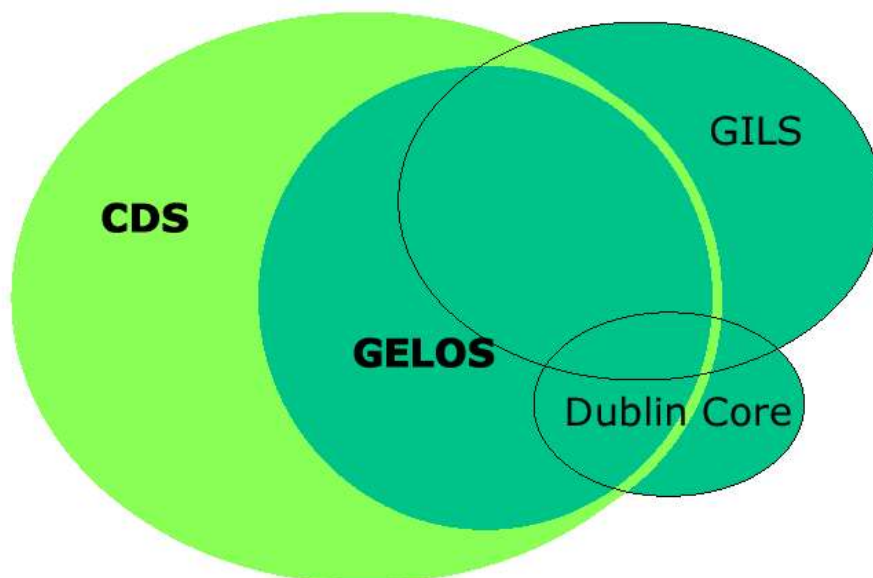


Figure 4 Overlapping metadata standards (Menger, Kazakos 2000)

XML provides an easy mechanism to define explicit namespaces for CDS, GELOS, and GILS (Dublin Core has a namespace⁸) and to integrate all of them in any XML document, as shown in an example in Figure 2.

3.1.2 Semantic Web

In the late 90s, environmental information providers started to use the internet as a publishing and communication platform. Today this has grown rapidly. For example, in the beginning of 1999, GEIN project found 50.000 Web pages of German environmental authorities. One year later, there were 80.000. In the view of metadata, this results in a completely new situation. The classical data catalogues are locator systems that deal with information that is hard to find and to access. So the information must be collected from lots of people who know what information exists and how to describe it. The description must be detailed enough for the user to decide, whether the specific data is of interest and if it is worth the effort to access it via individual communication by sending notes, papers and tapes or opening ftp connections. In the internet, information can be easily found by search engine crawlers, accessed immediately by hyperlinks. Metadata can be generated automatically using intelligent text analysis, and the end user just has to "klick" a hyperlink to have a first impression of the content itself.

That's why Web metadata can be gathered and kept up-to-date very easy. And that's why it needs not to be so detailed. When the GEIN metadata set was discussed in early 1999, we found out that only a core set of three attributes (topic, area, and time) is needed to supply a convenient navigation.

If the providers are willing to do so, they can place their own metadata in the headers of the information itself. This is the idea of the semantic web, as described by Tim Berners-Lee in the W3C DesignIssues⁹:

"The semantic web is a web of data, in some ways like a global database."

The most important standard in this context is the Resource Description Framework (RDF).

"RDF is a declarative language and provides a standard way for using XML to represent metadata in the form of statements about properties and relationships of items on the Web. Such items, known as resources, can be almost anything, provided it has a Web address. This means that you can associate metadata with a Web page, a graphic, an audio file, a movie clip, and so on."¹⁰

EML MetaData will build a large distributed catalogue of indexed web-based information. Each URL that is part of the environmental web, or the main entry points of a website, is described in terms of *topic*, *location* and *time*. This will be done based on a well-defined vocabulary, using

- GEMET, a multilingual environmental thesaurus, for the topic,
- United Nations Location Codes for the coding of the location, and
- international time formats conforming to ISO 8601.

EML MetaData should be modeled as an application of the Resource Description Framework (RDF).

3.2 EML Data eXchange

There are XML-based methods for the exchange of environmental data content. Each dataset to be exchanged can be described using predefined elements like:

- Core Attribute Types
- Generic Complex Attribute Types
- Predefined Document Types

These elements may be extended to fit to the needs of any given application. Any consumer can rely on a basic common understanding of the data structures that are exchanged.

3.3 Tools

The following tools must be developed and maintained:

- the GEMET vocabulary in the form of a Topic Map (ISO/IEC 13250)
- MetaData Generator
- Data eXchange definition support
- a broker for the navigation using the MetaData catalogue
- a generic renderer for the visualization of Data eXchange objects

3.4 EML'99 Initiative

These first proposals are based on the results of the 1st "Workshop on the Environmental Markup Language (EML)", which took place on November 25-26, 1999, at the Institute of Information Systems of Humboldt University in Berlin (<http://www.wiwi.hu-berlin.de/iwi/EML99/>). We would like to thank all participants for their contributions, including Mario Christ, Christian Dade, Daniel Görsch, Günther Hamann, Ralf Isenmann, Jens Großklags, Matthias Menger, Bernd Ritschel, and Erich Weihs.

EML'99 also started off a working group EML-EMIS (Environmental Management Information System) that is dedicated to a more detailed modeling of object classes for topics like "material and energy flow", "environmental account", "environmental performance indicator", "prevention

of pollution", "asset", or "document". The first results of this work are going to be published in the workshop proceedings¹¹.

4 Conclusions

In this paper we proposed a first design for an XML-based Environmental Markup Language. The proposal builds heavily on existing de facto standards for environmental metadata, such as CDS, UDK, and GEIN. EML serves to wrap and, wherever possible, to harmonize these standards. EML consists of two parts: EML MetaData and EML Data eXchange. EML MetaData is a special namespace and vocabulary used by information brokers that are part of the *Semantic Web* and have a special focus on environmental concerns. EML Data eXchange consists of a core set of attribute and document type definitions to be used when exchanging environmental data content.

- ¹ Thomas Bandholtz, Richard Bös and Maria Rüter: The German Environmental Information Network (GEIN). In: In Cremers, A.; Greve, K. (eds.): Computer Science for Environmental Protection '00. 2000.
- ² H.-A. Jacobsen/O. Günther/G. Riessen: (2000) Component Leasing on the World Wide Web. *Netnomics*. To be published.
- ³ see <http://www.umweltdatenkatalog.de/koudk/index.html>
- ⁴ Topic report No 5/1999. Catalogue of Data Sources. Annual topic update 1998. Prepared by Stefan Jensen. European Environmental Agency 1999
(<http://themes.eea.eu.int/toc.php/improvement/information?doc=39126&l=en>)
- ⁵ UNEP And Eea Cooperate To Develop A Global Catalogue Of Data Sources. UNEP Information Note 00/26. Nairobi 2000. And: A Common Global Vocabulary On The Environment. UNEP News Release 2000/5. Nairobi 2000.
- ⁶ Kerstin Grünefeld: Using XML for information exchange in CDS systems. The 7 th ETC/CDS Symposium on Catalogue of Data Sources and Thesaurus. Hannover 2000.
http://www.mu.niedersachsen.de/cds/etc-cds_neu/workshop_expo.html .
- ⁷ Matthias Menger, Wassilios Kazakos: Use of XML in the European Metainformation Locator System of ETC/CDS. In: Hans-Knud Arndt, Oliver Günther (2000).
- ⁸ Beckett, Dave, Eric Miller & Dan Brickley. Using Dublin Core in XML (2000-07-14). Working Draft. <http://purl.oclc.org/dc/documents/wd/dcmes-xml-20000714.htm>
- ⁹ Tim Berners-Lee: Semantic Web Roadmap. 1998
<http://www.w3.org/DesignIssues/Semantic.html>
- ¹⁰ W3C Metadata Activity Statement, 2000. <http://www.w3.org/Metadata/Activity.html>
- ¹¹ Hans-Knud Arndt, Oliver Günther (eds.): Environmental Markup Language (EML). Proceedings of Workshop 1, Berlin 1999. Metropolis. Marburg, Germany 2000.